Pure & Appl. Chem., Vol. 69, No. 11, pp. 2239–2243, 1997.
Printed in Great Britain.

Solvation and solubility of globular proteins

A. Ben-Naim

Department of Physical Chemistry The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Abstract: The solvation Gibbs energy of proteins is expressed in terms of its various ingredients (refs. 1,2). Estimation of the magnitude of these ingredients can lead to an estimate of the overall solvation Gibbs energy of the protein. As expected, the hydrophilic groups exposed to the solvent are mainly responsible for the solubility of the protein. However, we also found, unexpectedly, that correlation between hydrophilic groups on the surface of the protein can have a decisive contribution to the solvation Gibbs energy, and hence to the solubility of the protein.

1. INTRODUCTION

We begin with a short introduction to solvation thermodynamics (refs. 1,2). We define the solvation *process* as the process of transferring a solute from some fixed point in an ideal gas phase, into a fixed point in the liquid. The process is being carried out at constant pressure P, temperature T and composition N. Using the most common variables, in solution chemistry, namely P, T, N, we can write the chemical potential of the solute s in the ideal-gas, and in the liquid phases, as

$$\mu_s^l = \mu_s^{*l} + kT \ln \rho_s^l \Lambda_s^3 \tag{1}$$

$$\mu_s^g = \mu_s^{*g} + kT \ln \rho_s^g \Lambda_s^3 \tag{2}$$

where ρ_s^l and ρ_s^g are the number densities of s in the liquid and in the gaseous phases, respectively (note that these are not restricted to be small compared with the solvent density, as required in the conventional definition of the solvation Gibbs energies (ref. 1). Λ_s^3 is the momentum partition function, and it is assumed to be the same in the two phases. μ_s^* is referred to as the *pseudo-chemical potential* (PCP). It is defined in eq. (1) and eq. (2) for each phase. It should be noted that this quantity is different from the so called *excess chemical potential*. The latter is defined by

$$\mu_s^{Ex} = \mu_s^l - \mu_s^{i \cdot g} = \mu_s^l - [kT \ln \rho_s^g \Lambda_s^3 q_s^{-1}] \tag{3}$$

i.e., in Eq. (3) we extract the *ideal-gas* part of the chemical potential, which includes the term $kT \ln q_s$, where q_s is the internal partition function of the solute s. In eqs. (1) and (2), only the translational (or the liberation (ref. 1)) part of the chemical potential is extracted from μ_s^l.

Using statistical mechanical arguments it is easily shown that the Gibbs energy change for the solvation process as defined above is given by

$$\Delta G_s^*(g \rightarrow l) = \Delta \mu_s^*(g \rightarrow l) = \mu_s^{*l} - \mu_s^{*g} \tag{4}$$

For small solutes s, one can measure the solvation Gibbs energy (SGE) from the equilibrium partition of s between the two phases of l and g. This follows from the condition of equilibrium $\mu_s^l = \mu_s^g$, and from eqs. (1) and (2), i.e.

$$\Delta G_s^*(g \rightarrow l) = \mu_s^{*l} - \mu_s^{*g} = kT \ln(\rho_s^g / \rho_s^l)_{eq} \tag{5}$$

Clearly, for large solute molecules, such as proteins, having no measurable vapor pressure, the determination of ΔG_s^* through relation (5) is not possible. Nevertheless, since the knowledge of the solvation Gibbs energies of proteins is essential for estimating the driving forces of biochemical processes involving proteins, we appeal to theoretical estimates of these quantities. We shall next describe our procedure of estimating the solvation Gibbs energies of proteins.

2. BUILDING UP THE COMPONENTS OF THE SOLVATION GIBBS ENERGY OF A GLOBULAR PROTEIN

Proteins are large and very complicated molecules, having many different groups, each of which interact differently with water. Its structure is also not fixed in time. All these make an estimate of the Gibbs energy of solvation an almost impossible task. We describe here the procedure we adopt to make such an estimate, or at least to identify the most important factors that contribute to the Gibbs energy of solvation.

As our first step we can view the protein as a giant non-polar molecule. Clearly, it would be inappropriate to extrapolate the properties of such a large molecule of a diameter of about 50-80Å from the properties of small non-polar molecules with a diameter of about 3-5Å. A better way of viewing the protein, still as a giant non polar molecule, is as a cluster of, connected or disconnected, small molecules, closely packed to form a nearly spherical structure. Here, we can distinguish between two groups of molecules. Those that are exposed to the solvent, and those that are buried in the interior of the cluster, and therefore are not exposed to the solvent. Clearly, when the cluster becomes very large, the m_I particles in the interior (I) will determine the volume of the cluster, whereas the m_s particles on the surface(s) of the cluster will determine the surface area of the cluster. Accordingly, starting from the statistical mechanical expression for the solvation Gibbs energy (SGE) of a solute α (refs. 1,2), we write

$$\begin{aligned}\Delta G^*_\alpha &= -kT \ln < \exp[-\beta B_\alpha] >_o \\ &= -kT \ln < \exp[-\beta(B^H_\alpha + B^S_\alpha)] >_o \\ &= -kT \ln < \exp[-\beta B^H_\alpha] >_o - kT \ln < \exp[-\beta B^S_\alpha] >_H \\ &= \Delta G^{*H}_\alpha + \Delta G^{*S/H}_\alpha \end{aligned} \tag{6}$$

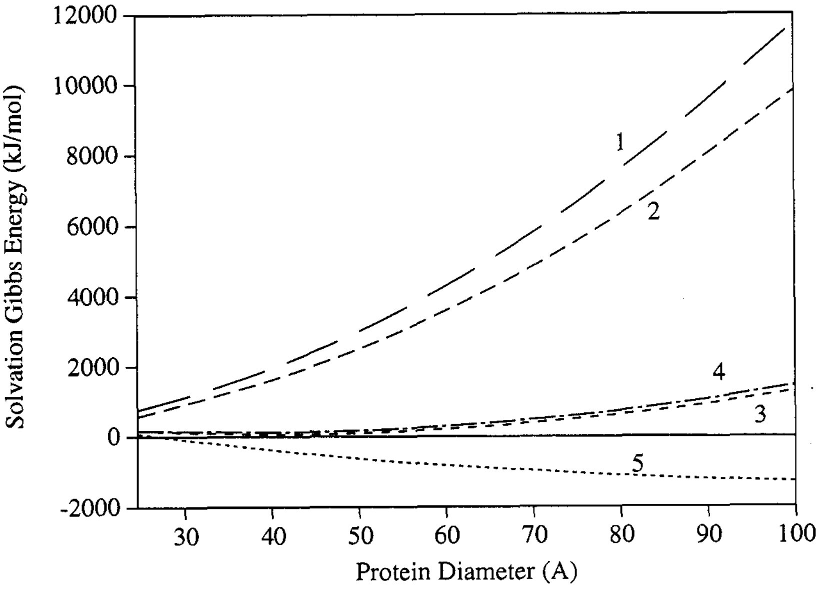

Fig. 1 Calculated curves of solvation Gibbs energy as a function of the protein diameter, using 5% as the criteria in determining whether a HΦI atom is exposed to the solvent. Curve 1 is the solvation Gibbs energy of cavity formation. In curve 2, the soft part of interaction between protein and the solvent is turned on. In curve 3, we further turn on the hydrogen bond part of interaction by treating all the HΦI atoms as if they were independently solvated. In curve 4, we add the negative correlations among HΦI atoms. Curve 5 is the total solvation Gibbs energy of protein, which results from the addition of positive correlations among HΦI atoms to curve 4.

In eq. (6) we have simply rewritten the SGE of α, our hypothetical cluster, in two terms. The fisrt is due to the hard (H) - repulsive interaction between α and its surrounding solvent molecules. It can be shown that this term is equivalent to the work required to create a cavity, at some fixed position in the solvent, capable of accommodating the entire solute α. The second term on the r.h.s. of eq. (6) is the additional work associated with "turning on" the soft (S) part, or the van der Waals interaction between the solute and the solvent. Once the cluster, representing our hypothetical protein is large enough, it is mainly the m_s particles that are exposed to the solvent that will interact with the solvent molecules through "soft" or van der Waals interaction. The cavity work is determined essentially by the total number of particles in the cluster.

ΔG_α^{*H} may be approximated by the scaled-particle-theory (ref. 3). Figure 1 shows the dependence of ΔG_α^{*H} on the diameter of the cavity. It is clear that ΔG_α^{*H} is everywhere positive and increases monotonically with the size of the protein. Had we extrapolated from ΔG_α^* of inert gas molecules we would have predicted negative values for ΔG_α^*. In Fig. 1 we also show the contribution of the van der Waals interaction between the m_s particles (on the surface) and the water molecules. Clearly, for large clusters of particles the relative contribution of the m_s - surface particles becomes negligible compared with the volume of the entire clusters. This means that for very large clusters the SGE will be large and positive. Thus, modeling of a protein by a giant non-polar molecule is inadequate to explain the high solubility of proteins. Such an explanation must be obtained by including the groups, on the surface of the protein, that "interact favorably" with water. We shall be more precise in defining the term "interact favorably" below. First, we add to our model, as depicted in Fig. 1, some side-chains that we know that they do interact favorably with the solvent, but we assume that these groups are far apart from each other, so that their solvation spheres do not overlap.

Since there are some 20 different side chains that could be exposed to the solvent, and since each of these could only be partially exposed to the solvent, we must deal with a very large number of possible groups - each of which interacts differently with the solvent. Therefore, in order to proceed we must make some drastic simplifications regarding the number of different groups that we are going to include in our analysis. We take only two representative groups, one referred to as hydrophobic (HΦO), which interacts unfavorably with water (see below) and the second, referred to as hydrophilic (HΦI), which interacts favorably with water. At this stage it is also assumed that these groups are far apart from each other.

We next address ourselves to the question of defining the term "interact favorably" with water. This is part of a more general question that has led to the construction of various "hydrophobicity scales". These scales assign numerical values to each of the amino acid, or to its side chain. These values are supposed to measure the extent of the affinity of the particular amino acid towards water. In my opinion, none of these scales is satisfactory, and as we shall point out below, there can be no such scale that will be completely satisfactory. To see this, let us first ask what are we looking for, in constructing a hydrophobicity scale. In the context of the solvation properties of protein, we seek a quantity that measures the relative contribution of each amino acid, or its side chain, to the solvation Gibbs energy of the entire protein. Unfortunately, there is no such quantity. The main reason for this is that the contribution of each side chain depends not only on the properties of that particular amino acid, but also on its neighbors, i.e., it depends on the environment in which that particular group is found. Clearly, the environment is different for different proteins.

There is, however, one case for which one can construct an "ideal" hydrophobicity scale. This occurs when the side chains on the surface of the protein are independently solvated (see below for a more precise definition, qualitatively we require here that the side chains that are exposed to the solvent be far apart from each other). In this case, eq. (6) is extended to

$$\Delta G_\alpha^* = \Delta G_\alpha^{*H} + \Delta G_\alpha^{*S/H} + \sum_{i=1}^{n} \Delta G_\alpha^{*i/S,H} \tag{7}$$

This expansion can be easily obtained from the statistical mechanical expression for the SGE. Qualitatively it means that in order to solvate the protein α, we can first "turn on" the hard part of the solute-solvent interaction. Next we "turn on" the soft part of the interaction, and finally we "turn on" each of the side-chains exposed to the solvent. The latter part produces the sum over i in eq. (7). This is valid only for the case when the n side chains are independently solvated. In this case, eq. (7) already suggests a definition of the required hydrophobicity scale. The quantities $\Delta G_\alpha^{*i/S,H}$ are referred to as the conditional SGE of the group i given that the hard (H) and soft (S) parts of the interaction have already been "turned on". If the groups are independently solvated then the quantities $\Delta G_\alpha^{*i/S,H}$ depend on the type of the backbone, i.e., the hard and soft interactions, but not on the other groups that are exposed to the solvent. We can now say that a group i interacts favorably or unfavorably with the solvent according to the value of $\Delta G_\alpha^{*i/S,H}$ of that group. The more negative the value of $\Delta G_\alpha^{*i/S,H}$, the larger will be the contribution of that group to the solubility of the protein. We stress that this assignment is meaningful only within the assumption of independence of the groups i. When the groups are not independently solvated the quantities $\Delta G^{*i/S,H}$ will in general depend on other groups in their vicinity. In such a case one cannot construct a hydrophobicity scale which assigns numerical values for each amino acid side-chain. To see this, and in order to obtain a precise definition of the concept of "independence" of the conditional SGE, consider the following simple example: Suppose that we have two side-chains, a and b exposed to the solvent, but which are close enough. Their contribution to the SGE of the entire protein is

$$\Delta G_{\alpha}^{*a,b/S,H} = -kT \ln < \exp[-\beta B_a - \beta B_b] >_{S,H} \tag{8}$$

where B_a and B_b are the total binding energies of a and b to the solvent, the average being taken over all the configurations of the solvent molecules, given that the hard and soft parts (S,H) of the interaction are already "turned on". We shall say that the two groups a and b are independently solvated, whenever this average can be factored into a product of two averages, namely

$$< \exp[-\beta B_a - \beta B_b] >_{SH} = < \exp[-\beta B_a] >_{S,H} < \exp[-\beta B_b] >_{S,H} \tag{9}$$

or, equivalently

$$\Delta G_{\alpha}^{*a,b/S,H} = \Delta G_{\alpha}^{*a/S,H} + \Delta G_{\alpha}^{*b/S,H} \tag{10}$$

Clearly, this occurs whenever the two groups are far apart (and therefore the two groups are uncorrelated in the probabilistic sense). Another way to write eq. (8) when the two groups are dependent is

$$\begin{aligned}\Delta G_{\alpha}^{*a,b/S,H} &= -kT \ln[< \exp[-\beta B_a] >_{S,H} < exp[-\beta B_b] >_{S,H,a}] \\ &= \Delta G_{\alpha}^{*a/S,H} + \Delta G_{\alpha}^{*b/S,H,a}\end{aligned} \tag{11}$$

Compare this eq. (11) with the previous one (eq. 10). Here we first turn-on the group a. The corresponding contribution to the SGE is $\Delta G_{\alpha}^{*a/S,H}$. Next, we turn on group b. Now, the corresponding contribution is not $\Delta G_{\alpha}^{*b/S,H}$ as in the case of independence (eq. 10), but it is a new conditional SGE, depending on S,H as well as on the presence of a. In view of the last equation, we can say that a and b are independently solvated when $\Delta G_{\alpha}^{*b/S,H,a}$ is independent of condition "a", i.e., when condition "a" can be eliminated without affecting the SGE of b.

It is now clear that if the side-chains, or functional groups that are exposed to the solvent are not independently solvated, there exists no hydrophobicity scale that will assign numerical values of hydrophobicity to the individual groups. The contribution of each group is strongly dependent on the other groups in its immediate vicinity. In the most general case, eq. (7) should be extended to read

$$\Delta G_{\alpha}^{*} = \Delta G_{\alpha}^{*H} + \Delta G_{\alpha}^{*S/H} + \sum_{i} \Delta G_{\alpha}^{*i/S,H} + \sum_{ij} \Delta G_{\alpha}^{*i,j/S,H} + \cdots \tag{12}$$

where the fourth term on the r.h.s. of eq. (12) includes all pair correlations between groups. Higher order correlations must be taken in an exact expansion of the SGE.

3. ESTIMATION OF THE SGE OF GLOBULAR PROTEINS

Having identified the various ingredients that contribute to the SGE of protein, we try to use all the information, theoretical or experimental, which is available to estimate the SGE of proteins. From the recently available data, we know that the conditional SGE of a HΦO group, such as methyl or ethyl, is positive and relatively small. On the other hand, the conditional SGE of a HΦI is negative and large, on the order of -7 kcal/mol. Therefore we shall focus only on those HΦI groups that are exposed to the solvent.

In the first stage of our estimate of the contribution of all the HΦI groups we assume that they are independently solvated. We then add the sum of all the contribution of these groups, using some criteria regarding the extent of exposure of each of these groups to the solvent. As expected, adding all the HΦI groups changes the SGE curve from highly positive to slightly negative. This is equivalent to an increase of solubility. As an example, a fully solvated group contributes about -7 kcal/mol to the SGE. This can be translated into an increase in solubility on the order of $\exp[-7/0.6] \sim 10^5$. This is a huge increment in the solubility per one (independently solvated) HΦI group.

Next, we add pair correlations between pairs of HΦI groups. With our limited information on these correlations we were obliged to use various simplifications to determine which pairs are positively and which are negatively correlated (ref. 4). Qualitatively we expect to find negative correlations whenever two HΦI groups are at a distance shorter than about 4Å. The reason for such a negative correlation is that the presence of one HΦI group, close to a second HΦI group may interface with the solvation of the second group. The possible loss of "favorable interaction" with the solvent, contributes positively to the SGE of the protein. On the other hand, when the two HΦI groups are at a range of distances between 4-5Å, the solvation of one group will enhance the solvation of the second group. This enhancement is due to the geometry of hydrogen-bonding by water molecules. If the two groups are ideally located at a distance of about 4-5Å, a solvating water molecule can form a hydrogen-bonded bridge between these two HΦI groups. It was estimated that this HΦI effect may contribute about -2.5 kcal/mol. These positive correlations will make the

SGE of the protein more negative. It is seen from Fig. 1 that the contribution of positive pair correlations is far larger (in absolute magnitude) than the negative correlations. We have estimated that the contribution of each (ideal) positive correlation to the total SGE is on the order of exp[2. 5/0. 6] ≈ 60.

From this analysis we can arrive at the following conclusion: A large non-polar molecule will have an extremely low solubility (this is mainly due to the extensive work required to create the suitable cavity). Adding the independently solvated HΦI group, brings the SGE of the protein to near-zero values, i.e., this eliminates much of the positive SGE due to the cavity work. Addition of pair correlations between HΦI groups gives rise to an additional negative contribution to the SGE. We suspect that these pair (and perhaps higher order) correlations are responsible for the large solubility of globular proteins.

REFERENCES

1. A. Ben-Naim, *Solvation Thermodynamics*, Plenum Press, New York (1987).
2. A. Ben-Naim, *Statistical Thermodynamics for Chemists and Biochemists*, Plenum Press, New York (1992).
3. H. Reiss, *Adv. Chem. Phys.* **9**, 1 (1966).
4. H. Wang and A. Ben-Naim, in preparation for publication.